Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 3 (2011) 1432–1438
Procedia Computer Science
www.elsevier.com/locate/procedia

WCIT 2010

Simultaneous feature selection and ant colony clustering

Emre Akarsu[a] *, Adem Karahoca[a]

[a]*Department of Software Engineering, Bahcesehir University, Istanbul, Besiktas, Turkey*

Abstract

Clustering is a widely studied problem in data mining. Ai techniques, evolutionary techniques and optimization techniques are applied to this field. In this study, a novel hybrid modeling approach proposed for clustering and feature selection. Ant colony clustering technique is used to segment breast cancer data set. To remove irrelevant or redundant features from data set for clustering Sequential Backward Search feature selection technique is applied. Feature selection and clustering algorithms are incorporated as a Wrapper. The results show that, the accuracy of the FS-ACO clustering approach is better than the filter approaches.

Keywords: Ant Colony Clustering; Simultaneous feature selection; SBS

1. Introduction

Clustering is grouping objects into a number of clusters according to their similarities in multidimensional spaces. Clustering is one of the challenging and important unsupervised learning techniques. There are a lot of applications of clustering in disciplines such as network partitioning [1], partitioning social network usage via user's patterns [2], in microbiology to identify the genome expressions [3] in computer vision image segmentation [4]. Meta-heuristic approaches are also applied to clustering problem [5]. In this study, Ant Colony Optimization (ACO) is used for clustering objects into similar groups.

Training a learning algorithm under high dimensionality domain is both computationally expensive and inefficient due to irrelevant or redundant feature's effect on empirical performance of the learning algorithm [6]. Also the probability of over fitting for the learning algorithm increases as the dimensionality of the training data set increase. Irrelevant or redundant features decrease the accuracy of the learning algorithm which is also known as the curse of the dimensionality [12]. In order to overcome these issues feature selection approach is applied. As the number of the features grows the complexity of the search space increases exponentially. For n features, every feature has the probability of being selected or not selected. So complexities of the feature search algorithms become 2^n.

Feature search algorithms classified into three categories as Complete, Sequential and Random search methods

1877-0509
doi:10.1016/j.procs.2011.01.026

with respect to Search strategy. Unsupervised feature selection is grouped into Filter and wrapper approaches [7]. Filters first select the features using an objective function independent from the learning algorithm then apply the clustering procedure. In wrappers, the feature selection is simultaneous with the clustering process. Wrappers make better performance of the clustering algorithm than the simple ranker methods [10].

In this study, a wrapper ant colony clustering and feature selection approach is proposed. This novel methodology is applied to the breast cancer data.

2. Breast Cancer dataset

Data set is gathered from UCI Machine Learning Repository. 286 samples are selected for training of the FS-ACO clustering algorithm. 9 features are selected for the segmentation of data. The attributes are categorical.

3. Methods

3.1 ACO for Clustering

Problem is partitioning N objects into K clusters such that the distances of the objects are minimized. R agents are used to construct solutions. Ants pick up objects based on the trail pheromone matrix.

$$p_{i,j} = \frac{\tau_{i,j}}{\sum_{i=1}^{k} \tau_{i,k}}, \mathrm{j} = 1..\mathrm{K} \tag{1}$$

The agents used the pheromone trail info to construct solutions. After each agent constructs their solutions, a local search is to get better solutions using the fitness values. Then the solutions with the highest fitness value are used for pheromone update [11].

The distance metric used in this study is different from the Shelokar's method [8]; instead of Euclidian distance we used Manhattan distance for similarity measurements between objects. The Manhattan distance is described as:

$$\left| x_{i,v} - m_{j,v} \right| \tag{2}$$

The objective of the clustering algorithm is the within cluster similarity which is known as the Scatter seperability criterion [9]. The fitness of the solutions is calculated by the Equation 2 above. It is the sum of the Manhattan distance between objects and the cluster centroids.

$$\text{Min F (w, m)} = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{v=1}^{n} w_{i,j} \left| x_{i,v} - m_{j,v} \right| \tag{3}$$

After selecting the top fitness value solutions, local search procedure is applied. Number of L new solutions is generated.

The third step of the ant lifecycle is the pheromone update process. After performing the local search and replacing the better fitness valued solutions, the pheromone updates implemented using these solutions. The fittest

solutions affect the pheromone trail matrix. L solutions are used to update pheromone trail matrix using the following pheromone updating formulation.

$$\tau_{i,j}(t+1) = (1-p)\tau_{i,j}(t) + \sum_{l=1}^{L} \Delta\tau_{i,j}^{l} \quad (4)$$

$i = 1,..,N, j = 1,..,K$

p： in this equation is the pheromone evaporation rate.

When pheromone evaporation parameter increases in exploration of new solutions becomes slowly, and if the evaporation parameter decreases it leads to forget the suboptimal solutions.

3.2 SBS Algorithm

Sequential backward selection starts the solution with full feature set and removes the feature with the worst fitness value from the feature subset. The search ends when no improvements can be done to the current feature set.

The algorithm starts with the full feature set. The function F is the evaluation function used for the learning algorithm. Then at each step one feature is removed, and after the remove operation the evaluation criterion is used to get the fitness value of the given subset. Among the solutions with one feature removed, the highest fitness value In the next iteration another feature is removed and the same process is repeated untill reaching the stopping criterion. Stopping criterion may be a priori to select the minimum subset of features or the search stops when no improvements can be achieved.

3.3 Feature Selecting Ant Colony Clustering (FS-ACO Clustering):

Wrapper approach has three integral parts. First one is the Feature Search algorithm which is the SBS algorithm in this study. The second part is the unsupervised learning algorithm which is the ACO clustering explained in the previous section. The final and the integral part of the algorithm is the evaluation function. In our approach, the unified criterion approach is used.

3.4 Unified Evaluation Criteria

In e Wrapper approach the feature selection and clustering processes become integrated. Scatter seperability index [7], will be used for clustering and feature search as unified criterion.

This index monotonically increases as the number of attributes increases. To be able to compare the fitness values of different feature sizes, a penalty term is used. Modified fitness function is:

$$trace(S_w^{-1}S_b) * (D - d + 1)/(D - 1) \quad (5)$$

D: is the number of attributes

d: is the number of selected attributes.

Each agent starts to build a solution. For initialization each ant randomly picks a cluster, the centroids of the clusters are computed and the evaluation criterion is firstly invoked to determine the performance of the iteration. After solution construction highest %1 percentage of fitness value of the solutions are selected for local search operation. After the local search if the new generated solution is better than the older one then new solution is replaced with the old one. The pheromone matrix is updated with only the top %l of the solutions. Then the sequential backward search procedure is invoked on the top solution. One feature is removed and performance of the solution is evaluated via evaluation function in iterations. Then if an improvement to the solution is done, the feature which's removal contributes the solution most is extracted from the solution. The process iterates until reaching the stopping criteria. Removal process stops when no further improvement can be assessed or the minimum number of the features is reached.

4. Findings

The results of this study are explained in this section. The proposed FS-ACO clustering method's performance evaluations on datasets are given.

Stopping criteria of the FS-ACO clustering algorithm is determines as the maximum number of iterations.

As the number of the agent increases the algorithm converges faster. As the number of feature to remove increases, to make the learning algorithm more stable the number of agents is increased.

In the first experiment, FS-ACO clustering algorithm removes 1 of the attributes from the training set. Parameters setting for this experiment are given below.

Table 1. FS-ACO Parameters

Parameter	Value	Parameter	Value
num_clusters	3	pheromone_evoporation	0.01
num_agents	30	local_search_threshold	0.01
default_phoeromone	0.001	local_search_num	6
pheromone_priori	0.1	max_iterations	3000

In order to compare the wrapper method with filter methods, K-means algorithm is used. To be able to select the features, full dataset is clustered via k-means then algorithm. With the labels of the clustering attribute selection is applied to the data set. To accomplish this task, Weka information gain feature selection is used. The results of the feature ranking are given below.

Table 2. Ranks of attributes for Weka clusters

# of attribute	rank	# of attribute	rank
10	0.3467	4	0.0943
2	0.2873	5	0.0795
3	0.2305	7	0.0755
8	0.1769	6	0.0559
9	0.1313		

The graphics in Figure 1 illustrates the sum of the error as the number of the removed features increases. Sum of error, which is the Manhattan distance between the objects and the centroids of the clusters, tends to reduce as the number of features decrease. FS-ACO feature selecting clustering algorithm outperforms Weka with information gain attribute selection.

Table 3. FS-ACO clustering vs. Weka sum of squared error comparison

# of features removed	Weka	Fs-Aco
1	732	758
2	575	731
3	433	658
4	319	435
5	239	293

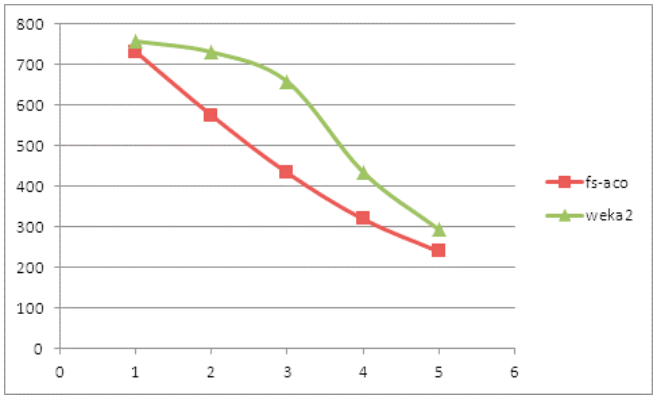

Figure 1. Comparison of the FS-ACO and Weka results

5. Conclusion

In this study, ACO clustering wrapping feature selection method is proposed. Algorithm designed using the stochastic nature of the ACO algorithm for clustering problem and combining it with the SBS methods for feature selection. Feature selection and the clustering algorithm are incorporated as wrapper. ACO method used in this study based on the Shelokar's Ant Colony Clustering model. Nominal attributes are used in this study so the distance metric for clustering is chosen as the Manhattan distance. Fitness function is improved by applying the CRIT criterion and a penalty term.

The novelty in this research is the incorporation of ACO clustering algorithm and the SBS method. In order to simultaneously select the features and partitioning the data items, the fittest solution is used. First iteration of the ACO clustering algorithm is done and then SBS algorithm is applied. The next iteration is generated using the selected features.

Finally to compare the results of the FS-ACO clustering algorithm, a filter approach is used. The data set with full feature set is clustered. The class labels are used for information gain ranking algorithm in Weka. Attributes are removed according to their ranks in the algorithm. FS-ACO clustering algorithm and the information gain feature selecting K-means algorithm's error rates are compared. FS-ACO algorithm outperformed K-means algorithm in sum of the squared errors and the selected feature set quality.

References

1. Bortner, D., & Han J., 2010 Progressive clustering of networks using Structure-Connected Order of Traversal Data Engineering (ICDE), IEEE 26th International Conference, 15 April 2010 Long Beach, CA pp. 653 – 656
2. Zhou Y., Fleischmann K., & Wallace W. 2010 Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors. Proceedings of the 43rd Hawaii International Conference on System Sciences 5-8 Jan. 2010 Honolulu, HI pp. 1 – 10
3. Dost B., Wu C., Su A. & Bafna V., 2010. TCLUST: A fast method for clustering genome-scale expression data. IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS 99, 20 May 2010 Available at http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5467032
4. Agrawal R., Grosky W.I., & Fotouhi, F., 2009 Searching an Appropriate Template Size for Multimodal Image Clustering. Multimedia Computing and Systems, ICMCS '09. 2-4 April 2009 Ouarzazate, pp. 560 – 564
5. Hruschka, E.R., Campello, R.J.G.B., Freitas A.A., & de Carvalho A.C.P.L.F., 2009 A Survey of Evolutionary Algorithms for Clustering. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, Vol. 9, Issue 2, March 2009 pp. 133 – 155
6. Gheyas I. A., & Smith L. S. Feature subset selection in large dimensionality domains. Pattern Recognition, Vol.43, Issue 1, January 2010, pp. 5-13
7. Dy J., & Brodley C. Feature Selection for Unsupervised Learning. Journal of Machine Learning Research(2004) pp. 845–889.
8. Shelokar P.S., Jayaraman V.K., & Kulkarni B.D.. An ant colony approach for clustering. Analytica Chimica Acta 509 (2004), pp. 187–195

9. Fukunaga K. Statistical Pattern Recognition. Academic Press, 1990.
10. Yay, M., & Akıncı, E. (2009). Application of Ordinal Logistic Regression and Artifical Neural Networks in a Study of Student Satistaction. Cypriot Journal Of Educational Sciences, 4(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/72.
11. Ozcinar, Z. (2009). Developing a Scale on the Instructional Communicatve Qualification of Parents With Teachers. Cypriot Journal Of Educational Sciences, 1(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/5.
12. Yılmaz, M., & Orhan, F. (2010). High school students educational usage of Internet and their learning approaches. World Journal On Educational Technology, 2(2). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/wjet/article/view/170